

Dancing Pigs or Externalities? Measuring the Rationality of Security Decisions

ELISSA M. REDMILES, MICHELLE L. MAZUREK, AND JOHN P. DICKERSON, University of Maryland

Accurately modeling human decision-making in security is critical to thinking about when, why, and how to recommend that users adopt certain secure behaviors. In this work, we conduct behavioral economics experiments to model the rationality of end-user security decision-making in a realistic online experimental system simulating a bank account. We ask participants to make a financially impactful security choice, in the face of transparent risks of account compromise and benefits offered by an optional security behavior (two-factor authentication). We measure the cost and utility of adopting the security behavior via measurements of time spent executing the behavior and estimates of the participant's wage. We find that more than 50% of our participants made rational (e.g., utility optimal) decisions, and we find that participants are more likely to behave rationally in the face of higher risk. Additionally, we find that users' decisions can be modeled well as a function of past behavior (anchoring effects), knowledge of costs, and to a lesser extent, users' awareness of risks and context ($R^2=0.61$). We also find evidence of endowment effects, as seen in other areas of economic and psychological decision-science literature, in our digital-security setting. Finally, using our data, we show theoretically that a "one-size-fits-all" emphasis on security can lead to market losses, but that adoption by a subset of users with higher risks or lower costs can lead to market gains.

1 INTRODUCTION

People's adoption, or rejection, of security behaviors can lead to system-wide consequences. While security requirements and system defaults such as automatic updates can remove many human decision-making fail-points for security-relevant systems, in many cases human decisions are still required to enable security. As such, security researchers have tried to understand and alter the decisions of the "human in the loop" [Cranor, 2008].

Prior work has proposed two simplified theories of the human in the loop: a rational actor who chooses to ignore security behaviors because the costs always outweigh the potential losses, and an irrational actor who chooses "dancing pigs over security every time" because they neither understand nor care about security risks [Herley, 2009]. While these simplified models of user behavior can help to provide high-level insights, our aim is to define a more realistic medium between these two extremes: *a semi (or boundedly) rational security actor with predictable, but not always utility-optimal, behavior based on risks and costs.*

End-user decision-making in security has largely been examined hypothetically [Böhme and Grossklags, 2011, Herley, 2009] or qualitatively [Felt et al., 2014, Harbach et al., 2013, Hong, 2012, Mathur et al., 2016, Sheng et al., 2010, Sunshine et al., 2009, Vaniea and Rashidi, 2016, Wash and Rader, 2015]. Here, we seek to provide an empirical, economic examination of the rationality of security behavior in a particular context. We define rationality as utility-optimality: that is, a decision is rational if the utility (gain) from the decision is greater than the costs of enacting the decision. Ultimately, we seek to understand: **How do costs (C), risks (R), and user tendencies and attributes (U) influence: (1) a security decision and (2) whether that decision is rational?**

By answering this question and quantifying a model of rationality in security behavior, we can guide users toward utility-maximizing behavior (or lack of behavior) in order to optimize the use of their personal, behavioral compliance budgets and maximize market gains from security [Acquisti and Grossklags, 2005, Beauteument et al., 2008].

Author's address: Elissa M. Redmiles, Michelle L. Mazurek, and John P. Dickerson, University of Maryland, [eredmiles, mmazurek, john]@cs.umd.edu.

In this work, we construct an experimental system and conduct behavioral-economics experiments (e.g., games) to evaluate and model security decision-making. Such experiments allow for a high level of control, realism, and quantification beyond that usually achieved from traditional survey and lab-based user-studies [Grossklags, 2007]. Our experimental system operates like a bank account to which users have to regularly log in. In each game, users are assigned to a condition: in one of the endowment conditions participants are given an amount of money and are required to login once every 24 hours to retain their money, while those in one of the earn conditions begin with a small amount of money and have the opportunity to earn more every time they log on. When signing up for a game, participants are offered a security choice: whether to enable *two-factor authentication* (2FA). Prior to making this choice they are shown explicit risks (which vary by game): risk of being hacked and amount of protection offered by adopting 2FA. If they are hacked – probabilistically determined by a script regularly run on the system – participants lose all of the money in their account. At the end of a game, participants receive the real monetary value of the amount left in their account.

Participants in our experiment play the game twice, each time with a different condition (endowment/earn, hack risk, and protection offered by 2FA). We invited 150 AMT workers to play the first game. Based on the observations we record during game play, we model the factors that influence whether a participant enables 2FA and whether the decision was utility optimal (e.g., rational). We further examine the extent to which people make utility-maximizing decisions and how the optimization of choices varies based on effective risk and the endowment vs. earn condition.

We find that people’s decisions to enable 2FA can be explained well (pseudo- $R^2 = 0.612$) by their prior behavior (approximately 35% of the variance), knowledge of costs (e.g., awareness of how long 2FA will take) (approx. 15% of behavior variance), and to a lesser but still significant extent, explicit risk judgements and endowment effects (9% to 15% of variance, depending on their knowledge of other factors). Further, we find that both risk and endowment effects also relate to whether the decision to enable a behavior was utility-optimal. We find that those playing in the lowest-risk conditions are far less likely to make utility-optimal decisions than those in the highest-risk conditions. Those in the medium-risk conditions make a utility-optimal decision 58% of the time the first time they play the game and 69% of the time the second. Our results suggest that, in our experiment:

- Users made rational security choices approximately 50% of the time.
- Users tend to act in accordance with an anchoring effect [Strack and Mussweiler, 1997]: they tend to stick with the first security decision they make. More general security tendencies (chosen password strength, security behavioral intentions) did not influence behavior.
- Users are boundedly rational: they appear to incorporate knowledge about costs and explicit risks.
- In higher-risk conditions, users enable security options more often and make rational decisions more often.
- When protecting assets they already have, especially in highest-risk games, users tend to behave more rationally and more securely (endowment effect [Knetsch, 1989]).

Using this data, we can compute estimates of utility-optimizing behaviors for a given user based on users’ typical costs (e.g., login speed, account balance, and earning potential) and estimated explicit risks (risk of hacking, protection offered by a behavior). As a tangible example, we present a thought experiment using our data to theorize about market gains and losses from different approaches to user security: encouraging all users to be secure, encouraging utility-optimal behavior, and not requiring end users to engage in security behaviors. We show that a “one-size-fits-all” emphasis on security can lead to significant market losses from needless user costs for negligible

security utility, and identify hypothetical risk thresholds at which targeted encouragement of certain security behaviors for at-risk users can lead to market gains.

Consequently, we suggest a push toward personalized security recommendations (e.g., nudges) to guide users toward utility-optimal behaviors and allow them more autonomy to make decisions based on transparently communicated risks paired with a push for more data-driven research quantifying those risks. We conclude with further discussion of the implications of our findings for predicting user behaviors and improving market, or population, security through nudging and risk communication, and suggest directions for future work.

2 RELATED WORK

Below, we review prior work on the decision-making in security and privacy, focusing specifically on economics-based studies when available.

Security Decision-Making. Herley takes the first economic approach to examining end-user security decisions [Herley, 2009]. In his theoretical paper, he provides a cost-benefit analysis of having end-users follow security advice. He argues that threats are so rare – and suggested behaviors so ineffective – as to make it logical for end users to never adopt security behaviors. Relatedly, Böhme and Grossklags approach user attention in security as a public good, that is, as an extremely scarce resource and examine how to facilitate optimizing rather than satisficing of user attention toward security warnings and privacy notices [Böhme and Grossklags, 2011]. They theoretically explore the effect of different user interface designs on the use of attention and recommend that user interactions should be rationed, rather than just user interfaces optimized, to maximize the use of users’ privacy and security compliance budgets.

The majority of security-related economics analysis of human behavior has focused on experts’ (e.g., system or network administrators) responses to threat [Christin et al., 2004, Grossklags et al., 2008b, Hota and Sundaram, 2015, Johnson et al., 2010], organizational responses to threat [Campbell et al., 2003, Gal-Or and Ghose, 2005, Miura-Ko et al., 2008], or attackers [Alpcan and Basar, 2006, Chen and Leneutre, 2009, Liu et al., 2006, Manshaei et al., 2013, Sallhammar et al., 2005, Shim et al., 2012]. For example, Grossklags and colleagues conduct a number of different studies applying game-theoretic models to examine expert behavior in attack/defense security games [Christin et al., 2004, Grossklags et al., 2008a,b, Johnson et al., 2010] finding evidence of unclear risks and bounded rationality for security professionals; and relatedly, Rounds et al. examined the impact of changes in security and reward on attacker responses, finding that as the value of an account increases, so too does the number of attacks on that account; and as the amount of security on an account increases so too does the number of attacks decrease [Rounds et al., 2013]. These analyses of experts, however, provide limited insight into end-user behavior.

Christin et al. present one of the few empirical, economic-related, studies of security behavior: they incentivized users to execute arbitrary code [Christin et al., 2011]. They adjusted the incentives provided, finding that 43% of participants were willing to run untrusted executables for \$1.00. Most prior work focused on end-user decision making has been more qualitative, examining people’s security decisions through surveys and interview studies. These studies include explorations of why people say they update or do not update their software [Mathur et al., 2016, Vaniea and Rashidi, 2016, Wash and Rader, 2015], why users fall for phishing attacks [Hong, 2012, Sheng et al., 2010], why people bypass browser warning messages [Felt et al., 2014, Harbach et al., 2013, Sunshine et al., 2009], and even preliminary support for boundedly-rational decision-making in cybersecurity based on post-event surveys with university students [Aytes and Connolly, 2004]. Findings have shown evidence of the importance of fatigue from too many security demands, unclear risks, and lack of clarity on the utility of behaviors for addressing risks.

The reasonable conclusion of many of these end-user studies is a push for better education of users to get them to take security risks seriously and enable security behaviors. Indeed, especially in the corporate realm, there may often exist an implicit assumption that more security is better and our goal should be to get all users to enable all security behaviors at all times [Herley, 2017, Kirsch and Boss, 2007]. Herley’s work challenges this approach, arguing that security is *never* better for the end-user, and at best benefits the corporations of whom they are clients while costing the individual [Herley, 2009]. Underlying both approaches is also an implicit assumption that users have rational reasons for adopting or ignoring security advice, and are able to interpret and make reasonable decisions about risks if they are aware of them. In our work, we explore to what extent users are rational actors and we explore what factors, including implicit and explicit risks and costs, relate to the security decisions they make in our simulation. We also explore the impact of the “users should always be secure” and “users should never be secure” perspectives on the costs and utility of enabling 2FA for an entire population to better understand when, and whether, we should encourage end users to adopt security behaviors.

Economics of Privacy. Prior work on privacy has explored privacy as a good and theoretical privacy markets [Acquisti et al., 2016, Berthold and Böhme, 2010, Farrell, 2012, Hui and Png, 2006] as well as empirical examinations of users valuations of data [Bauer et al., 2012, Carrascal et al., 2013, Cvrcek et al., 2006, Friedman, 2007], and especially their willingness to pay to protect private, digital information vs. willingness to disclose information for a payout [Acquisti et al., 2013, Schreiner and Hess, 2013, Schudy and Utikal, 2017]. Acquisti and colleagues have found evidence of bounded rationality – that is, semi-rational decisions in which people are only able to take into account a portion of the data (i.e., risks or evidence) in their decisions [Acquisti, 2004, Acquisti and Grossklags, 2005] and have found that difficulty identifying risks and benefits can lead to significant uncertainty that muddies privacy decision-making [Acquisti et al., 2015]. Closely related to our work, Tsai et al. explore what happens when making privacy tradeoffs more transparent, finding that when privacy information is surfaced to users in an online shopping scenario they are more likely to make purchases from a privacy-preserving website, even at a higher cost [Tsai et al., 2011]. Beyond privacy, behavioral economics experiments similar to ours have been used effectively to model bounded-ly rational decisions in business and management settings, health, and purchasing [Diamond and Vartiainen, 2012, Hanoch et al., 2007, Simon, 1979].

Our work draws on findings from these studies to explore bounded-rationality in security and examine how security-tendencies such as password strength choices influence decision-making. We combine the individual-focused analyses presented by this prior work in privacy with the market-level analyses emphasized by security economics work, presenting a perspective on how and when we should emphasize security protection for end-users.

3 METHODOLOGY

Below we describe our experimental system, the experiments (e.g., games), participant recruitment, our analyses, and limitations. All procedures were approved by our institution’s review board.

3.1 System Design

We created the “bank” system to provide a similar UI flow to logging into a real bank account (Figure 1 provides a system overview). The system URL was `bank.[institutiondomain]`, had an SSL certificate, and was thus displayed over HTTPS.

When participants began the study, they were taken to a page that had two options: log in or sign up. After choosing the signup button they were taken to a page where they could create an account. The account creation page asked the participant for their Amazon Mechanical Turk (MTurk) ID and to create a password. There were no constraints for password creation, and the

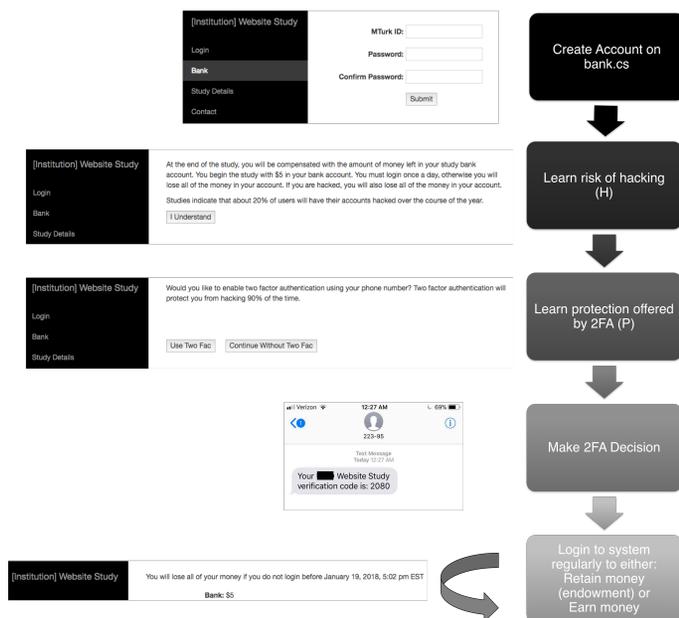


Fig. 1. Overview of the experimental system flow.

provided MTurk IDs were validated against the list of workers who accepted the task on MTurk. If the participant clicked the Login button, they would be taken to a page asking for their MTurk ID and their previously created password.

The next screen explained the study. In the study, we varied participants’ game condition, which was a combination of experimental setting (Earn or Endowment), risk of hacking (H), and protection offered by 2FA (P). Those in the Earn Setting ($S : Earn$) started with \$1 in their account and could earn an extra \$1 per day up to a maximum of \$5. Those in Endowment Setting ($S : Endowment$) started with \$5 and would lose that money if they did not log in once per day. In addition to varying the conditions, we also transparently told participants their risk of getting hacked H . Hacking was experimentally constructed: we ran a script every evening, randomly selected a number, and “hacked” the participant (e.g., removed the money from their account) if the randomly selected number was less than their effective risk; hacking probabilities were not affected by how frequently participants chose to login. The participant’s effective risk was defined as either H (no 2FA selected) or $H * (1 - P)$ (2FA selected), that is, the probability of being hacked modulated by the accepted protection. In the initial consent form and study instructions it was made clear to participants that the hacking was constructed and that there were not real hackers trying to break into their study accounts.

Participants were randomly assigned one of three H values: 20% (a chance of hacking popularly cited, which we use as a benchmark [Norton, 2012, Seymour, 2016, Taylor, 2016]), 1%, and 50%. To convey this information, participants in $S : Earn$ saw the message shown in the second screen from the top in Figure 1. Those in $S : Endowment$ saw a very similar message, which replaced the second and third sentences in the message with: “You begin the study with \$1 in your bank account. Each time you login (at most once per day) you will earn an additional \$1.”

Next, participants were offered the option to enable 2FA. We chose 2FA as the exemplar behavior in our system because prior work finds that it is an “in-the-middle” behavior – understood

better and adopted more often than updating, but less so than strong passwords [Redmiles et al., 2016]. Additionally, it was easier to create a simple, realistic system around a binary 2FA choice (enable/don't enable) than around other security behaviors such as updating or browser warnings.

We were explicit about the amount of protection offered by 2FA, just as we were explicit about the risk of being hacked. Participants were randomly assigned an arbitrarily chosen protection value, P , of 50% or 90%. As defined above, that value was then used in the effective risk calculation that led to a chance of “hacking” in our experiment. Thus, their effective risk was $H(1 - P)$. To communicate this, participants saw the following message: “Would you like to enable two factor authentication using your phone number? Two factor authentication will protect you from hacking P of the time.”

Those who selected 2FA went through SMS-based two-factor setup, where they received an initial text with a four-digit code that they needed to enter before proceeding.¹ Then, on all subsequent logins, they received an SMS with a code and had to enter this code along with their password in order to log in.

Finally, participants were shown their bank balance, and also reminded of when they must next log in, for $S : \textit{Endowment}$, or when they could next earn more money, for $S : \textit{Earn}$.

Upon subsequent logins, participants would see the previously described login page where they logged in using their MTurk ID, password, and, if they selected the option, 2FA. After logging in, the participant was directed to their bank balance page. At the end of the study, on their last login, they were shown a link to a Qualtrics survey which the user needed to complete in order to receive payment. The survey used two validated measures, SeBIS [Egelman and Peer, 2015] and the Web-Use Skills Index [Hargittai and Hsieh, 2012], to measure the participant’s general security behavior intentions and internet skill; we also collected their gender, age, and education level.

Technical Details: Data Storage & Measurement. Our system was built using PHP 5.6.32. The 2FA validation was handled by the AuthyAPI. All data was stored in a secure database. Only the hashed and salted password, not the cleartext password, was saved using the PHP password-hash function with PASSWORD_DEFAULT. Passwords were verified using the PHP password-verify function.

To measure password strength, we used the data-driven neural-network password meter [Ur et al., 2017]. To record the signup and login times, we recorded the total number of seconds for which the 2FA screen was in focus. We began recording when the page was loaded and stopped recording focus time when the pages were submitted. That is, if a user went to the login page, stayed on that page for five seconds, checked their email for two minutes, and finally came back to the website and took another ten seconds to submit their login, we would have recorded a login screen time of fifteen seconds. Table 5 in Appendix 11 provides a summary of all observed and self-report collected variables.

3.2 Experimental Design

We recruited 150 workers from Amazon Mechanical Turk (MTurk) to participate in two rounds of an experiment run on our bank system. We recruited only workers from the United States and those who had a 95% approval rating or above, as prior work shows that these workers produce the highest quality results [Peer et al., 2014]. Our advertisement stated that participants would be evaluating a website for our institution and had the opportunity to earn up to \$5. In the first experiment,

¹While SMS-based 2FA has been shown to be insecure [Reaves et al., 2016], it is still the most broadly adopted 2FA standard. Additionally, SMS-based 2FA provides a simple, quantifiable privacy and time cost for our experimental scenario: app-based 2FA would complicate cost calculations, as we would not know if certain users already had the application we chose installed on their phones. For purposes of our experiment, the actual security offered by SMS-based 2FA in practice does not matter.

participants completed a consent form and were then redirected to create an account on the bank site. They progressed through the steps described above, and were assigned to one of eight experimental settings: H=1%, P=50%, Endowment and Earn (Low Risk); H=1%, P=90%, Endowment; H=20%, P=50% Endowment and Earn; H=50% and P=50% Endowment; H=50% and P=90% Endowment and Earn (High Risk). We selected a non-full-factorial design due to cost constraints; we selected the following conditions that would maximize our ability to draw conclusions about the effect of various risk levels as well as the endowment condition. The experiment ran for five days. All participants who started the experiment were paid \$0.50, and those who submitted their final surveys were paid the amount remaining in their bank account (ranging from \$0 through \$5.00).

Five days after the conclusion of Round 1 (RD1), participants received an email through MTurk inviting them to participate in the experiment again. They were told that things would work the same way as last time, but that they needed to create a new account. Participants were again randomly assigned to one of the eight experimental settings, with no relation to their RD1 settings. When participants finished Round 2 (RD2) they were not asked to take the survey again, but were again compensated as in RD1. Participants spent an average of a total of 142 seconds (Std. Dev = 34.57 (s)) logging in to RD1 and an average of a total of 158 seconds (Std. Dev = 30.40 (s)) logging into RD2, plus an average of 6 minutes completing the survey after the first game.

3.3 Analysis

Our analysis consists of three components: (1) modeling security behavior as a function of the variables measured and assigned in the games; (2) determining the degree to which participants' behavior was utility-optimal, that is "rational," and modeling the factors correlating with rational decision-making; and (3) presenting a thought experiment showing how we could use this information to predict utility-optimal behaviors for MTurker's and how such predictions could be used to nudge security behavior in the future. We describe our specific methodology for each analysis in more detail in Sections 5 to 7. Table 6 in the Appendix summarizes the models we construct.

3.4 Limitations

Behavioral economics experiments are subject to a number of limitations: participants may behave differently than in real life, our variables for the hack and protect percentages may be unrealistic, and 2FA may not be a representative security behavior. We have done our best to mitigate these limitations by setting our incentive in line with typical behavior economics incentives [Kamenica, 2012] and by choosing hack and protection percentages both higher and lower than real values suggested by popular studies for the hack percentage, and we tried to pick protection values that would provide information about whether people feel differently about a nearly "perfectly" protective security behavior and a behavior that will only work half of the time. Finally, we chose 2FA as our exemplar behavior because prior work suggests it is a reasonable "middle-of-the-road" in terms of user understanding and adoption [Redmiles et al., 2016].

It is also possible that the source of our sample, Amazon Mechanical Turk, and the fact that our sample is not representative of the demographics of the US may have biased our results (our participants are younger and more educated than the U.S. population, as is typical of AMT samples [Ross et al., 2010]); additionally it is possible that those participants who lost their money due to hacking or due to forgetting to log in would be far less likely to return their surveys or return for the second round. We chose to use Mechanical Turk because census-representative survey platforms do not allow for the deployment of experiments such as ours, and we experienced low drop out rates (17% RD1, 15% RD2) suggesting that lack of compensation for participants who were hacked or forgot to log in was not a significant source of bias.

4 PARTICIPANT DESCRIPTIVES

125 of the original 150 participants completed RD1. 107 of these RD1 participants finished RD2. The demographics of the two groups were nearly identical. Our samples were skewed slightly male (54% in R1, 55% in R2). Our participants were more educated than the general population, with 51% and 50% in each round, respectively, holding a college degree or above as compared to 28% of the U.S. population [U.S. Census Bureau, 2014]. Finally, our samples are both skewed heavily younger than the general population, with 70% of our participants under 40 years old compared to 40% of the general population. Table 4 in the Appendix shows the demographics of our samples.

Our RD1 participants had a mean internet skill score of 4.06 on the Web-Use Skills Index [Hargittai and Hsieh, 2012]; our RD2 participants had a mean internet skill score of 4.016 (this difference was not significant by a Mann-Whitney U test). Our RD1 participants had a mean SeBIS score [Egelman and Peer, 2015] of 3.17; RD2 participants had a mean SeBIS score of 3.16 (the difference in SeBIS scores across rounds was not significant by a Mann-Whitney U test).

5 WHAT FACTORS DRIVE PEOPLE'S SECURITY DECISION-MAKING?

Analysis Approach. To model our participants' behavior, we constructed binomial logistic regression models where enabling 2FA was the outcome variable. For modeling 2FA enablement in RD1, we included the following input variables: the participant's gender, age, educational level, internet skill, security behavior intention and password strength, as well as the protect and hack percentages presented, the setting (earn or endow), and two-way interactions between the latter three factors. We used backward selection based on the Bayesian Information Criterion to select the model of best fit [Spiegelhalter et al., 2002]. We report model fit using McFadden's Pseudo- R^2 [Cameron and Windmeijer, 1997], approximating the percent of variance explained.

For modeling round two behavior, we construct three models: one with only RD1 behavior as the input variable; one with RD1 behavior, RD1 signup time (not including time spent on 2FA), and RD1 mean login time (not including time spent on 2FA); and one with RD1 behavior and times as well as RD2 equivalents of the factors found to be relevant in the RD1 2FA enablement model (a subset of: hack and protection percentages, demographics, password strength, the setting, and relevant interactions) as the input variables. We chose to construct these models in order to examine the relationship between security decision-making and anchoring effects,² participants' costs (e.g., tendency to login or signup slowly), and the relevant factors identified in RD1 (our "blind" experiment, in which we do not yet know participants' typical tendencies or costs). We do not include whether the participant was hacked or lost money in RD1 because fewer than 10 such participants returned to play in RD2, a quantity insufficient to provide robust regression results. We compared the fits of our three models using likelihood ratio tests [Bentler and Bonett, 1980].

We report the results of all models, including log-adjusted regression coefficients (odds ratios), 95% confidence intervals for those odds ratios, p-values (with a significance threshold of 0.05), pseudo- R^2 , and the results of model comparison tests, when appropriate.

Factors that Drive 2FA Decisions. 51% of participants in RD1 and 56% in RD2 chose to enable 2FA. These decisions varied by condition, as shown in Figure 2.

We model RD1 decisions to enable 2FA as a function of attributes of the respondent (gender, age, education, internet skill), and security practices or savviness (security behavior intention, password strength), and risks and conditions assigned to the participant (H , P , S), as well as interactions between these risks and settings: interaction between endowment (T/F) and H and interaction between endowment and P . We find that the model of best fit retains only the risk, protection, and

²The anchoring effect is a commonly shown psychological effect in which humans tend to stick with the first decision they make or choice they are offered [Strack and Mussweiler, 1997].

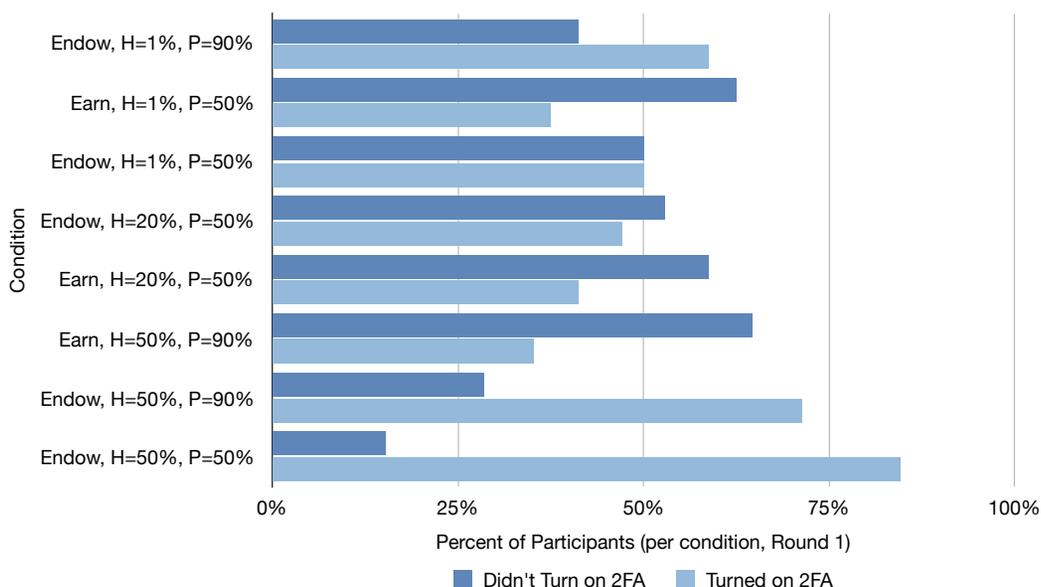


Fig. 2. Percent of participants who enabled 2FA in each of our eight experimental conditions in RD1.

Variable	O.R.	C.I.	p-value
<i>Endowment</i>	2.32	[1.44, 3.76]	<0.001*
<i>H</i>	2.31	[1.22, 4.38]	0.011*
<i>P</i>	1.46	[1.22, 1.97]	0.043*
<i>Endowment:P</i>	3.61	[1.35, 9.67]	0.012*

Table 1. Logistic regression results for 2FA behavior in RD1. O.R. is the log-adjusted regression coefficient (odds ratio), CI is the 95% confidence interval of that odds ratio, and p-values marked with * are significant at the 0.05 level. Endowment is a boolean variable for whether the participant was in one of the endowment conditions, *H* and *P* are numeric, and ':' indicates an interaction variable.

setting factors, as well as the interaction between setting and protect percentage (Table 1). Those in the endowment conditions are 2.3 times more likely to enable 2FA, in line with endowment effects observed in other fields [Acquisti et al., 2013, Knetsch, 1989].

Those who are shown a higher *H*, that is told that research shows a higher chance of account hacking, are more likely to enable 2FA. Those shown a higher *P*, that is, those told 2FA will be 90% rather than 50% effective, are also more likely to enable 2FA, while those in a condition that involved endowment and a higher *P* are even more more likely to enable than just accounting for *P* or the endowment setting separately would suggest. This model explains 14.7% of the variance in the RD1 data.³

As described above, we model RD2 decisions with three different sets of models. The model with just RD1 behavior as an input factor indicates that those who enabled 2FA in RD1 are 83 times

³We note that this is likely a *conservative* interpretation of the effect; indeed, psychologists have suggested that R^2 values may appear deceptively low when modeling singular events (e.g., decisions) that tend to accumulate over time, and that the true explanatory power of the significant variables in these cases may be stronger than a singular measurement indicates [Abelson, 1985].

Variable	O.R.	C.I.	p-value
R1 2FA Decision	32.66	[2.19, 486.54]	0.011*
R1 Login Times	0.81	[0.6, 1.11]	0.190
R1 Signup Time	0.90	[0.69, 1.16]	0.411
<i>Endowment</i>	1.26	[0, 547.12]	0.094
<i>P</i>	17.09	[0.04, 144.74]	0.357
<i>H</i>	72.88	[0.57, 373.07]	0.063
<i>Endowment:P</i>	0.99	[0.91, 1.08]	0.822
R1 2FA:Login Times	1.12	[0.81, 1.56]	0.496
R1 2FA:Signup Time	2.41	[0.26, 22.02]	0.435

Table 2. Logistic regression results for 2FA behavior in RD2 modeled as a function of RD1 2FA decision, RD1 costs (login and signup times), and RD2 experimental settings. McFadden Pseudo- $R^2=0.612$

more likely to do so in RD2 (Appendix Table 7), in line with the expected anchoring effect. This model explains 35.3% of the variance in our RD2 data. Next, we added RD1 costs (e.g., RD1 signup and login times) to the model to understand whether RD2 behavior was explained just by repeating RD1 behavior or also by learning from the costs of RD1 behavior. The model of best fit retained all three factors, and we see that both login time in RD1 and RD1 behavior are significant in the model, with those who enabled in RD1 being more likely to enable in RD2 and those who took longer than others to login in RD1 being 6% less likely to enable 2FA in RD2 (Appendix Table 8). This model explains the data significantly better than the prior model ($p = 0.012$) with pseudo- $R^2=0.522$. Finally, we built a third model that included these three RD1 factors as well as the experimental settings retained in RD1. This final model shows only that enabling 2FA in RD1 is significantly related to enabling 2FA in RD2, however, we find that this final model with the experimental variables fits significantly better than the prior model ($p = 0.040$) and explains a total of 61.2% of the variance (Table 2), suggesting that the costs and experimental settings are explanatory of behavior but not as strongly as are anchoring effects.

6 DO PEOPLE MAKE RATIONAL SECURITY DECISIONS?

Analysis Approach. To examine whether participants made utility-optimal decisions and the factors related to the rationality of those decisions, we considered participants to have made a *utility-optimal* decision in the following way: it is utility-optimal to enable 2FA if the cost of doing so is less than the expected utility that would be gained from 2FA. For those who enabled 2FA in either round, we compute the cost of using 2FA for an individual user as the time it cost them to sign up (in hours) plus the sum of the time it cost them to log in each time (in hours) times their hourly wage: $C_{2fa} = (T_{signup} + \sum T_{login}) * wage_{mturk}$. To compute the value of our participants' *wage* we used the data collected in the Pew 2016 survey about MTurk workers [Hitlin, 2016] to compute a population-based estimate of \$4.97/hr (calculations in Appendix 12). For those who did not use 2FA, and in any analysis in which we use the full dataset (that is the data of those who enabled and did not enable 2FA), we estimate the cost of 2FA. A sensitivity analysis reveals a linear pattern in estimations: considering 2FA cost as the mean results in an estimate of more rational choices than considering 2FA cost as $2\times$ or $3\times$ the mean. To generate conservative estimates of rational behavior we continue our computations by estimating 2FA cost for those who did not enable as $2\times$ the mean.

We compute the utility of using 2FA as the potential loss (maximum amount they could earn times hack percentage) times the protection gained by using 2FA (P): $U_{2fa} = P[(H) * Max_{bank}]$. Where Max_{bank} was \$5 for our experiment, and is the value of the money in an MTurker's account for the third analysis.

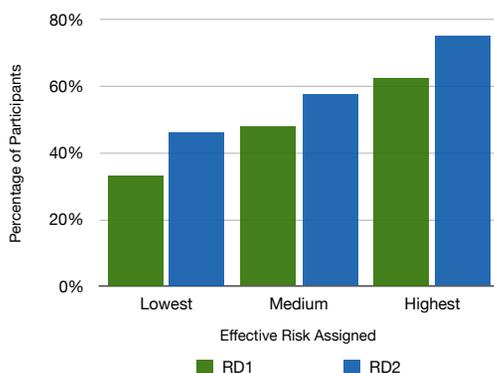


Fig. 3. Percentage of participants in each round and with a given risk setting who made a utility-optimal decision about 2FA.

Below we examine whether 2FA users made utility-optimal decisions. We also examine the proportion of utility-optimal decisions made by our sample as a whole using the mean cost of enabling 2FA (computed from those who did choose to use it). To compare the strength of the associations between the decisions made in each round, we use a χ^2 test with Cramér’s V as our measure of effect size and use Holm-Bonferroni correction to account for multiple comparisons.

Finally, we model whether participants’ decisions were utility-rational using regression models. We construct these models using a similar approach to that described in Section 5, with RD1 utility-optimal decision-making being modeled as a function of the participant’s gender, age, educational level, internet skill, security behavior intention and password strength, as well as the protect and hack percentages presented, the setting, and two-way interactions among the latter three factors. RD2 utility-optimal decision-making (True or False) was modeled with two models: the first was a single-factor model containing only whether the RD1 decision was utility-optimal, and the second included the same factors as in RD1 as well as the RD1 decision. We performed backward selection on this second model to achieve a model of best fit for RD2.

Were Participants’ Decisions Rational? We find that, following the approach described above, 48% of all participants made utility-optimal decisions in RD1 and 58% did so in RD2.

We also consider behavior at the extremes: that is, the lowest and highest risk conditions, as well as in the middle. We find that, in RD1, 33% of participants in the lowest-risk settings (e.g., did not enable 2FA), 48% in the medium risk, and 63% in the highest risk settings make a utility-optimal decision (e.g., enabled 2FA⁴). We observe a learning effect ($\chi^2 = 21.226$, $df = 2$, $p < 0.001$, $V = 0.578$ (medium)) with 58% of all participants in the medium-risk experiments making a rational decision in RD2, 46% in the lowest-risk settings and 75% in the highest-risk settings. Figure 3 compares the change in utility-optimal behavior by risk level for the rounds.

Modeling utility-optimal decision making. We wanted to understand how these different factors, as well as participant factors, interacted to potentially explain whether participants’ make utility-optimal decisions.

We first model whether participants made a utility-optimal 2FA decision in RD1 in the same way as we modeled general decision-making for RD1 in Section 5 above. The model of best fit retains only the hack percentage, setting, and internet skill factors (Appendix Table 9; pseudo- $R^2 = 0.141$). Those in the endowment setting are 25% more likely to make a utility-optimal choice, while those

⁴In our particular experiment, using 2FA was always utility-optimal in the highest risk setting and never in the lowest.

with higher internet skill are 15% more likely to do so. Further, we see that those who saw a higher hack percentage are more likely to make a *utility-optimal decision* in RD1.

For RD2, we find that a single-factor consisting of only whether the RD1 decision was utility optimal is not significant. After backward selection, the multi-factor model of best fit retains SeBIS, the risk and protection factors, endowment, and interactions between risk and endowment and protection and endowment (Appendix Table 10; pseudo- $R^2=0.078$). We see that those with higher SeBIS scores are nearly four times more likely to have made a utility-optimal decision in RD2, and (similar to RD1) those who saw a higher hack percentage are 21% more likely to make a utility-optimal decision.

7 THOUGHT EXPERIMENT: WHEN DOES SECURITY BENEFIT THE MARKET?

In order to examine the impact of different models of user behavior on market costs and utility from security, we present a simple thought experiment regarding the implementation of optional 2FA on the MTurk platform.

Approach. Suppose that MTurk decided to offer a 2FA option to workers. Workers who enable 2FA then must use 2FA every time they log in, exactly as in our experiment. We know from prior work [Hitlin, 2016] that 63% of MTurkers log in daily, 32% multiple times a week (for the purposes of our thought experiment, we say 3× per week), 3% log in weekly, and the rest log in less than once a week. Using costs (as defined in Section 6) from our experiment, this means 2FA would cost: $C_{2fa} = 0.065 * wage_{mturk} + 0.039 * N(logins) * wage_{mturk}$. Using this equation, we see that each week using 2FA would cost the daily login group \$0.34, the 3× a week login group \$0.18, and the weekly login group \$0.10.

Suppose also that MTurkers transfer the money they earn out of their accounts once per week, so the value of their bank (e.g., the money they stand to lose if their account is hacked) is the value of their weekly earnings. Not all MTurkers earn the same amount on MTurk; prior work shows that 17% earn all of their income on MTurk, 8% earn most (75%), 6% earn half (50%), 15% earn some (25%), and 53% earn very little (5%) of their income on the platform. We estimate, based on reports from the population that those who earn all or most of their income on MTurk make \$38,555 per year or \$741 per week in total, with either 100% or 75% of that income coming from MTurk. Those who earn half or less of their income on MTurk earn a higher overall income of \$54,300 or \$1044 per week, but only 5%-50% of that income comes from MTurk. The header of Table 11 shows the estimated weekly earnings – that is, the amount of money an MTurker would stand to lose – for the different categories of earnings.

Finally, we estimate the utility gained from 2FA for some people at an arbitrarily selected 1% risk of hacking over the course of the year and for others at an arbitrary 20% risk of hacking over the course of the year. We suppose in this thought experiment that 25% of people might choose weak passwords, thus having a hypothetically supposed 20% risk of hacking over the course of the year. The remaining 75% of the population we theorize have a 1% risk of hacking. Thus, the utility of 2FA in a given week would be calculated as follows (with a conservatively assumed 50% protection value): $U_{2fa} = P(H * Income)/52$. We again consider enabling 2FA to be utility optimal in cases where the utility is greater than the cost.

It is important to note that we choose fairly arbitrary values for this thought experiment, so the results should be considered as an example rather than taken literally. However, we believe this example (as with prior theoretical work that uses a similar approach [Herley, 2009]) provides a useful perspective for thinking carefully about when and how to advise users.

Results. We find that under the above assumptions, for all but those who earn the least on MTurk (\$52 per week) those who are at a 20% risk of hacking and who log in more than once a week should always enable 2FA. Those that log in once a week could enable 2FA or not enable 2FA

Approach	Costs	Utility	Loss/Gain
All Secure	\$275	\$148	(-) \$126
Utility Optimal	\$32	\$128	(+) \$96
No Security	\$0	\$0	\$0*

Table 3. Estimated costs, utility, and loss or gain from each of the security approaches. Figures are per 1000 MTurk users. *We use the [Herley, 2009] model, in which users sustain no losses from hacking, to compute this value. If users sustained losses, the value would instead be \$266.

with no discrepancy in outcome, as the cost of enabling and the utility gain are equal. For those who earn the least on MTurk, it is never utility optimal to enable 2FA.

Based on these calculations, we compute the cost and utility, and the overall gain or loss, from the three different recommendations for user security: users should always take all security precautions, users should never take security precautions, and users should take utility-optimal security precautions. Table 3 summarizes the results per 1000 MTurkers. We see that utility-optimal behavior leads to a gain of \$96 per 1000 MTurkers, while an “enable security always” solution leads to a cost of \$126 per 1000 MTurkers, and a “never enable security” solution leads to a gain of \$0. Estimates suggest that there are 500,000 MTurkers active at any given time [Hitlin, 2016]. Using this estimate, requiring all MTurkers to use 2FA would cost the market \$63,606, while providing MTurkers with nudges for suggested behavior that is utility-optimal would lead to a gain of \$47,865 if all MTurkers behaved optimally and \$23,932 if 50% chose a utility optimal behavior. Of course, the current option of not presenting 2FA leads to no gain or loss.

More generally, assuming that 2FA offers protection only 50% of the time, we find that for those who have \$741 in their accounts (e.g., earn \$741 weekly and withdraw their income once per week) there need only be a 1.5% chance of their account being hacked over the course of the year for there to be utility gain from 2FA. For those who login weekly, a 2.7% risk for those who login three times a week, and a 4.9% risk for those who login daily. Similarly, for those who earn half their income from MTurk (\$522) and login weekly there need only be a 2.2% chance of their account getting hacked over the course of the year for there to be utility gain from 2FA. (These figures include only monetary losses, not time losses or emotional consequences from cleaning up from hacking; including these costs would cause 2FA to become utility optimal at even lower risks.) On the other hand, for those who earn very little (or keep very little money in their MTurk accounts), the risks must be very large – 68% for daily logins or 21% for weekly logins – for 2FA to be utility optimal. Table 11 in the Appendix summarizes different risk and income levels at which 2FA may be utility optimal.

8 DISCUSSION

Make Risks, Costs, and Benefits Transparent. Overall, we find that explicitly stated risks and utilities (e.g., benefits) appear to explain between 9% (Round 2) and 15% (Round 1) of the variance in our participants’ behavior. While participants lean heavily on anchoring effects (e.g., their typical practices) and to a lesser extent their knowledge of their own costs (e.g., how long it takes them to perform a login) when settling on a behavior, the additional contribution of risk and utility variables to the variance explained suggests that *numeric* descriptions of risk and benefit are at least partially informing users’ security decision making. This finding agrees with prior findings in privacy that showed people were able to pick between different platforms verbally described to be more or less privacy protective [Tsai et al., 2011]. These findings consequently offer support for making risk and benefit information more transparent to users. Rather than telling people they

“should adopt behavior X to be more secure” we may wish to give them more autonomy over their choices.

Inherently, making risks (or benefits from a given behavior) more transparent requires us to know what the risk or benefit of a particular attack or behavior is to the user. Measuring this risk effectively is a challenging open problem [Herley, 2017]. Recent work, such as that conducted by Bilge et al. and Thomas et al. [Bilge et al., 2017, Thomas et al., 2017], pushes for more scientific and measurement oriented approaches to assessing the potential for an attack or the benefit of a security behavior. Our results provide additional motivation for this research direction, and present a parallel direction for future work exploring how best to convey risk to users (e.g., exploring how they interpret and respond to different textual and numeric estimates of risk), as well as investigating how explicit risk communication may factor into existing legal and policy frameworks around digital security and privacy disclosure.

Recommend Utility-Optimal Behaviors. In our experiment, being able to compute a cost for an individual’s security behavior and their expected utility from that behavior allows us to predict what users should do in a particular security decision-making scenario. A similar principle could be used in a real-world system to develop personalized security recommendations to guide users toward utility-optimal behaviors, rather than always trying to encourage adoption of security behaviors. That is, by measuring users’ typical time to complete certain behaviors (e.g., login), their implicit risk (e.g., password strength, security settings), and the history of attacks on a given platform for users with a similar risk profile, systems may be able to offer users personalized, suggested behaviors.

For example, a platform could display a message similar to the following on login or when the user is confronted with a security decision: “Our personalized security analysis system suggests that two-factor authentication may be a good fit for securing your account. We expect to take only an additional 30 seconds of your time.” While prior work has found that encouraging users to adopt behaviors that their friends have adopted [Das et al., 2014] can succeed, doing so may not always be of benefit to the user. A more carefully personalized system that takes into account the utility-cost ratio for the user, as well as other factors unexplored here, such as risk aversion, may offer a more trustworthy set of suggestions that are respectful of user time and reduce their security burden.

Security-Savvy Not Necessarily Safer. Our findings reveal that security-savvy, measured by password strength and SeBIS score, had no impact on choice to enable 2FA, and an inconsistent impact on rationality of decision. We had anticipated a relationship between 2FA use and password strength (or security behavior intention), either that stronger passwords (or more security intention) would indicate higher paranoia that would correlate with more 2FA use, or that those who picked weaker passwords would be more likely to enable 2FA, suggesting a tradeoff between the two risk-reducing behaviors. Instead, we find only a relationship between explicitly stated risks (e.g., the chance of getting hacked that we conveyed to them). This suggests that users treat different security behaviors independently without considering them as a package that influences overall risk or may be unaware of the relationship between their own choices (e.g., the strength of the password they pick) and their risks. If so, they may have difficulty making rational choices.

Thus, future work may wish to explore users’ level of understanding of the risks of their security choices, how they think these risks interact, and investigate how to correct resultant misperceptions. Additionally, we may wish to understand the efficacy of translating user choices into explicit risk assessments; for example, instead of presenting a scale of password strength from weak to strong, present password strength in terms of risk of account compromise.

Future Work. The work presented here takes a first step toward measuring the rationality of user decision-making, assessing the contexts in which more and less rational behavior occurs, and understanding how risks, costs, and user factors influence decisions in security. Our experimental

scenario presents the simplest of situations: with explicit, monetary, easily characterizable costs and risks. Future work may wish to explore rationality of decision making around data or other non-monetary assets and the impact of non-numeric risk statements, or explore other cognitive biases, e.g., present bias [Gravin et al., 2016, Kleinberg et al., 2016, 2017]. Learning and/or comparing against other models of bounded rationality is of practical interest [Hartford et al., 2016, Wright and Leyton-Brown, 2014]. Additionally, our experiment only explores users' choice to enable 2FA, and indirectly their choice of password strength. Future work should validate, or invalidate, these results with other security behaviors.

Finally, our results show that self-report measures of security behavior intention and internet skill were not related to 2FA enablement behavior. However, we see evidence that skill and/or intent may be related to the rationality of decision-making. This suggests that self-report measures may be ineffective at predicting or assessing behaviors — a finding supported by extensive work in other fields [Fendrich and Vaughn, 1994, Redmiles et al., 2017, Tourangeau and Yan, 2007] — but may be useful in assessing more deeply underlying constructs that are related to rationality. Future work should seek to further explore the utility of self-report measures for assessing security behavior, and whether experimental systems such as that presented here more, less, or equally as useful as self-reporting for measuring real-world behavior.

9 CONCLUSION

In this work we explore the rationality of decisions made by end-users when faced with security choices. We validate that, at least in our relatively realistic experimental scenario, end-users behave in accordance with bounded rationality. That is, they are able to take into account some, but not all, risk and context factors and make rational (e.g., utility optimal) decisions in more than 50% of cases. We find, further, that we are able to explain user decision making well (McFadden Pseudo- $R^2=0.61$) by accounting for participants' knowledge of costs and risks as well as accounting for the known phenomena of endowment and anchoring effects.

We find that our participants made utility-optimal decisions more often when faced with higher risks. While perhaps encouraging for corporate high-risk scenarios, this finding also suggests a challenge for day-to-day security, as many of the risks end users confront in daily digital life are less transparent, less monetarily linked, and relatively small. Thus, future work may wish to explore how rationality is affected by different methods of communicating risk, less tangible consequences than the monetary incentives provided in our experiments, and even smaller risks.

Using our data to compute utilities and costs for 2FA on the Amazon Mechanical Turk (MTurk) platform, we show that a “one-size-fits-all” approach encouraging all users to engage in all security behaviors at all times can lead to significant market and individual losses. However, we also show that there are relatively low risk thresholds at which 2FA becomes utility optimal, leading to significant market gains. This underscores the importance of quantifying the risk and utility of security behaviors through data-driven research [Bilge et al., 2017, Thomas et al., 2017] in order to ensure that risks, costs, and benefits are accurately understood.

Overall, our work supports a nuanced model of the “human-in-the-loop” who is able to somewhat take into account explicit risks and personal costs, in addition to relying on typical tendencies (e.g., anchoring effects), to make frequently rational decisions but who struggles to identify less obvious risks such as those incurred from weak passwords. This argues for personalized security-behavior recommendations for users tailored based on their costs (e.g., login times), risks (e.g., password strength), and value of their account (e.g., measured through the amount of money stored), especially in cases where the optimal behavior for them goes against their typical tendencies. If accurate, such personalized recommendations could provide security benefits while helping to avoid large market and personal costs from wasted time and effort on non-utility-optimizing security behaviors.

ACKNOWLEDGEMENTS

Elissa Redmiles acknowledges support from the National Science Foundation Graduate Research Fellowship Program under Grant No. DGE 1322106 and a Facebook Fellowship.

REFERENCES

- Robert P Abelson. 1985. A variance explanation paradox: When a little is a lot. *Psychological Bulletin* 97, 1 (1985), 129.
- Alessandro Acquisti. 2004. Privacy in electronic commerce and the economics of immediate gratification. In *Proceedings of the ACM Conference on Electronic Commerce (EC)*. ACM, 21–29.
- Alessandro Acquisti, Laura Brandimarte, and George Loewenstein. 2015. Privacy and human behavior in the age of information. *Science* (2015).
- Alessandro Acquisti and Jens Grossklags. 2005. Privacy and rationality in individual decision making. *IEEE Security & Privacy* 3, 1 (2005).
- Alessandro Acquisti, Leslie K John, and George Loewenstein. 2013. What is privacy worth? *The Journal of Legal Studies* 42, 2 (2013).
- Alessandro Acquisti, Curtis R Taylor, and Liad Wagman. 2016. The economics of privacy. (2016).
- Tansu Alpcan and Tamer Basar. 2006. An intrusion detection game with limited observations. In *12th Int. Symp. on Dynamic Games and Applications, Sophia Antipolis, France*, Vol. 26.
- Kregg Aytes and Terry Connolly. 2004. Computer security and risky computing practices: A rational choice perspective. *JOEUC* 16, 3 (2004).
- Christine Bauer, Jana Korunovska, and Sarah Spiekermann. 2012. On the value of information-what Facebook users are willing to pay. *ECIS* (2012).
- A. Beaument, M. A. Sasse, and M. Wonham. 2008. The compliance budget: Managing security behaviour in organisations. In *workshop on new security paradigms*. DOI : <http://dx.doi.org/10.1145/1595676.1595684>
- Peter M Bentler and Douglas G Bonett. 1980. Significance tests and goodness of fit in the analysis of covariance structures. *Psychological bulletin* 88, 3 (1980).
- Stefan Berthold and Rainer Böhme. 2010. Valuating privacy with option pricing theory. *Economics of information security and privacy* (2010).
- Leyla Bilge, Yufei Han, and Matteo Dell’Amico. 2017. RiskTeller: Predicting the Risk of Cyber Incidents. In *CCS*.
- Rainer Böhme and Jens Grossklags. 2011. The security cost of cheap user interaction. In *NSPW*. ACM.
- A Colin Cameron and Frank AG Windmeijer. 1997. An R-squared measure of goodness of fit for some common nonlinear regression models. *Journal of Econometrics* 77, 2 (1997).
- Katherine Campbell, Lawrence A Gordon, Martin P Loeb, and Lei Zhou. 2003. The economic cost of publicly announced information security breaches: empirical evidence from the stock market. *Journal of Computer Security* 11, 3 (2003).
- Juan Pablo Carrascal, Christopher Riederer, Vijay Erramilli, Mauro Cherubini, and Rodrigo de Oliveira. 2013. Your browsing behavior for a big mac: Economics of personal information online. In *WWW*. ACM.
- Lin Chen and Jean Leneutre. 2009. A game theoretical framework on intrusion detection in heterogeneous networks. *IEEE Transactions on Information Forensics and Security* 4, 2 (2009).
- Nicolas Christin, Serge Egelman, Timothy Vidas, and Jens Grossklags. 2011. It’s all about the Benjamins: An empirical study on incentivizing users to ignore security advice. In *FC*.
- Nicolas Christin, Jens Grossklags, and John Chuang. 2004. Near rationality and competitive equilibria in networked systems. In *SIGCOMM workshop on Practice and theory of incentives in networked systems*. ACM.
- Lorrie Faith Cranor. 2008. A framework for reasoning about the human in the loop. *UPSEC* 8, 2008 (2008).
- Dan Cvrcek, Marek Kumpost, Vashek Matyas, and George Danezis. 2006. A study on the value of location privacy. In *Proceedings of the 5th ACM Workshop on Privacy in Electronic Society*. ACM.
- S. Das, A. D.I. Kramer, L. A. Dabbish, and J. I. Hong. 2014. Increasing Security Sensitivity With Social Proof: A Large-Scale Experimental Confirmation. In *CCS*. DOI : <http://dx.doi.org/10.1145/2660267.2660271>
- Peter Diamond and Hannu Vartiainen. 2012. *Behavioral economics and its applications*. Princeton University Press.
- S. Egelman and E. Peer. 2015. Scaling the Security Wall: Developing a Security Behavior Intentions Scale (SeBIS). In *CHI*.
- Joseph Farrell. 2012. Can privacy be just another good. *Journal on Telecomm. & High Tech* 10 (2012).
- Adrienne Porter Felt, Robert W Reeder, Hazim Almuhammedi, and Sunny Consolvo. 2014. Experimenting at scale with google chrome’s SSL warning. In *CHI*. ACM.
- Michael Fendrich and Connie M Vaughn. 1994. Diminished lifetime substance use over time: An inquiry into differential underreporting. *Public Opinion Quarterly* 58, 1 (1994).
- Allan Friedman. 2007. An Empirical Approach to Understanding Privacy Valuation. (2007).
- Esther Gal-Or and Anindya Ghose. 2005. The economic incentives for sharing security information. *Information Systems Research* 16, 2 (2005).

- Nick Gravin, Nicole Immorlica, Brendan Lucier, and Emmanouil Pountourakis. 2016. Procrastination with variable present bias. *EC* (2016).
- Jens Grossklags. 2007. Experimental economics and experimental computer science: A survey. In *Workshop on Experimental computer science*. ACM.
- Jens Grossklags, Nicolas Christin, and John Chuang. 2008a. Secure or insure?: a game-theoretic analysis of information security games. In *WWW*. ACM.
- Jens Grossklags, Nicolas Christin, and John Chuang. 2008b. Security Investment (Failures) in Five Economic Environments: A Comparison of Homogeneous and Heterogeneous User Agents.. In *WEIS*.
- Yaniv Hanoch, Stacey Wood, and Thomas Rice. 2007. Bounded rationality, emotions and older adult decision making: Not so fast and yet so frugal. *Human Development* 50, 6 (2007).
- Marian Harbach, Sascha Fahl, Polina Yakovleva, and Matthew Smith. 2013. Sorry, I don't get it: An analysis of warning message texts. In *FC*.
- Eszter Hargittai and Yuli Patrick Hsieh. 2012. Succinct survey measures of web-use skills. *Social Science Computer Review* 30, 1 (2012), 95–107.
- Jason S Hartford, James R Wright, and Kevin Leyton-Brown. 2016. Deep learning for predicting human strategic behavior. In *NIPS*. 2424–2432.
- C. Herley. 2009. So Long, and No Thanks for the Externalities: The Rational Rejection of Security Advice by Users. In *NPSW*. DOI: <http://dx.doi.org/10.1145/1719030.1719050>
- Cormac Herley. 2017. Justifying Security Measures: a Position Paper. In *ESoRCS*.
- Paul Hitlin. 2016. Research in the Crowdsourcing Age, a Case Study. (2016).
- Jason Hong. 2012. The state of phishing attacks. *Commun. ACM* 55, 1 (2012).
- Ashish R Hota and Shreyas Sundaram. 2015. Interdependent security games under behavioral probability weighting. In *International Conference on Decision and Game Theory for Security*.
- Kai-Lung Hui and Ivan PL Png. 2006. Economics of privacy. (2006).
- Benjamin Johnson, Jens Grossklags, Nicolas Christin, and John Chuang. 2010. Are Security Experts Useful? Bayesian Nash Equilibria for Network Security Games with Limited Information.. In *ESORICS*.
- Emir Kamenica. 2012. Behavioral economics and psychology of incentives. *Annu. Rev. Econ.* (2012).
- Laurie Kirsch and Scott Boss. 2007. The last line of defense: motivating employees to follow corporate security guidelines. *ICIS* (2007).
- Jon Kleinberg, Sigal Oren, and Manish Raghavan. 2016. Planning problems for sophisticated agents with present bias. In *EC*. ACM, 343–360.
- Jon Kleinberg, Sigal Oren, and Manish Raghavan. 2017. Planning with multiple biases. In *EC*. ACM, 567–584.
- Jack L Knetsch. 1989. The endowment effect and evidence of nonreversible indifference curves. *American Economic review* 79, 5 (1989).
- Yu Liu, Cristina Comaniciu, and Hong Man. 2006. A Bayesian game approach for intrusion detection in wireless ad hoc networks. In *Proceeding from the 2006 workshop on Game theory for communications and networks*. ACM.
- Mohammad Hossein Manshaei, Quanyan Zhu, Tansu Alpcan, Tamer Bacşar, and Jean-Pierre Hubaux. 2013. Game theory meets network security and privacy. *CSUR* 45, 3 (2013).
- Arunesh Mathur, Josefine Engel, Sonam Sobti, Victoria Chang, and Marshini Chetty. 2016. " They Keep Coming Back Like Zombies": Improving Software Updating Interfaces.. In *SOUPS*.
- R Ann Miura-Ko, Benjamin Yolken, John Mitchell, and Nicholas Bambos. 2008. Security decision-making among interdependent organizations. In *CSF*. IEEE.
- Norton. 2012. Norton 2012 Cybercrime Report. (2012). https://now-static.norton.com/now/en/pu/images/Promotions/2012/cybercrimeReport/2012_Norton_Cybercrime_Report_Master_FINAL_050912.pdf
- Eyal Peer, Joachim Vosgerau, and Alessandro Acquisti. 2014. Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior research methods* 46, 4 (2014).
- Bradley Reaves, Nolen Scaife, Dave Tian, Logan Blue, Patrick Traynor, and Kevin RB Butler. 2016. Sending out an SMS: Characterizing the Security of the SMS Ecosystem with Public Gateways. In *Security and Privacy (SP), 2016 IEEE Symposium on*. IEEE, 339–356.
- Elissa M Redmiles, Yasemin Acar, Sascha Fahl, and Michelle L Mazurek. 2017. *A Summary of Survey Methodology Best Practices for Security and Privacy Researchers*. Technical Report.
- Elissa M Redmiles, Sean Kross, and Michelle L Mazurek. 2016. How i learned to be secure: a census-representative survey of security advice sources and behavior. In *CCS*. ACM.
- Joel Ross, Lilly Irani, M Silberman, Andrew Zaldivar, and Bill Tomlinson. 2010. Who are the crowdworkers?: shifting demographics in mechanical turk. In *CHI'10 extended abstracts on Human factors in computing systems*. ACM, 2863–2872.
- Mark Rounds, Norman Pendegraft, and Jim Alves-Foss. 2013. An Experimental Study to Explore Attacker Response to Changes in Security and Reward. In *HICSS*. IEEE.

- Karin Sallhammar, Svein J Knapskog, and Bjarne E Helvik. 2005. Using stochastic game theory to compute the expected behavior of attackers. In *Applications and the Internet Workshops, 2005. Saint Workshops 2005. The 2005 Symposium on. IEEE*.
- Michel Schreiner and Thomas Hess. 2013. On The Willingness To Pay For Privacy As A Freemium Model: First Empirical Evidence.. In *ECIS*.
- Simeon Schudy and Verena Utikal. 2017. "You must not know about me" On the willingness to share personal data. *Journal of Economic Behavior & Organization* (2017).
- John Seymour. 2016. How Often Are Social Media Accounts Hacked? (2016). <https://www.zerofox.com/blog/often-social-media-accounts-hacked/>
- Steve Sheng, Mandy Holbrook, Ponnurangam Kumaraguru, Lorrie Faith Cranor, and Julie Downs. 2010. Who falls for phish?: a demographic analysis of phishing susceptibility and effectiveness of interventions. In *CHI*. ACM.
- Woohyun Shim, Luca Allodi, and Fabio Massacci. 2012. Crime pays if you are just an average hacker. In *CyberSecurity*. IEEE.
- Herbert A Simon. 1979. Rational decision making in business organizations. *American Economic Review* 69, 4 (1979).
- David J Spiegelhalter, Nicola G Best, Bradley P Carlin, and Angelika Van Der Linde. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64, 4 (2002).
- Fritz Strack and Thomas Mussweiler. 1997. Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *Journal of personality and social psychology* 73, 3 (1997).
- Joshua Sunshine, Serge Egelman, Hazim Almuhiemedi, Neha Atri, and Lorrie Faith Cranor. 2009. Crying Wolf: An Empirical Study of SSL Warning Effectiveness.. In *USENIX security*.
- Ben Taylor. 2016. Why There is a 1 in 3 Chance You'll get Hacked in 2016. (2016). <https://www.bestvpn.com/privacy-news/get-hacked-one-in-three/>
- Kurt Thomas, Frank Li, Ali Zand, Jacob Barrett, Juri Ranieri, Luca Invernizzi, Yarik Markov, Oxana Comanescu, Vijay Eranti, Angelika Moscicki, Daniel Margolis, Vern Paxson, and Elie Bursztein (Eds.). 2017. *Data breaches, phishing, or malware? Understanding the risks of stolen credentials*.
- Roger Tourangeau and Ting Yan. 2007. Sensitive questions in surveys. *Psychological bulletin* 133, 5 (2007).
- Janice Y Tsai, Serge Egelman, Lorrie Cranor, and Alessandro Acquisti. 2011. The effect of online privacy information on purchasing behavior: An experimental study. *Inform*s (2011).
- Blase Ur, Felicia Alfieri, Maung Aung, Lujo Bauer, Nicolas Christin, Jessica Colnago, Lorrie Faith Cranor, Henry Dixon, Pardis Emami Naeini, Hana Habib, and others. 2017. Design and evaluation of a data-driven password meter. In *CHI*. ACM.
- U.S. Census Bureau. 2014. American Community Survey 5-Year Estimates. (2014).
- Kami Vaniea and Yasmeen Rashidi. 2016. Tales of software updates: The process of updating software. In *CHI*. ACM.
- Rick Wash and Emilee J Rader. 2015. Too Much Knowledge? Security Beliefs and Protective Behaviors Among United States Internet Users.. In *SOUPS*.
- James R Wright and Kevin Leyton-Brown. 2014. Level-0 meta-models for predicting human behavior in games. In *EC*. ACM, 857–874.

APPENDIX

10 PARTICIPANT DEMOGRAPHICS

Table 4 describes the demographics of the MTurk users who participated in both rounds of our experiment. The table compares their demographics to those of the U.S. Census [U.S. Census Bureau, 2014]. Note that our participants are more educated (51% and 50% of the participants in the rounds, respectively, hold a B.S. compared to 28% of the U.S. population) and younger (74% and 73%, respectively, are under the age of 39 compared to 40% of the U.S. population).

Metric	Round 1	Round 2	Census
Male	54%	55%	49%
Female	45%	44%	51%
HS or Less	13%	13%	28%
Some College	36%	37%	31%
B.S. or Above	51%	50%	28%
18-29 years	31%	32%	23%
30-39 years	43%	41%	17%
40-49 years	15%	15%	17%
50-59 years	8%	9%	18%
60+ years	2%	2%	25%

Table 4. Demographics of participants in our sample. Some percentages may not add to 100% due to item non-response. Census statistics from the American Community Survey [U.S. Census Bureau, 2014].

11 VARIABLES MEASURED IN THE EXPERIMENT

In Table 5 we summarize the variables measured in our experiment and subsequently included in our regression analyses. We outline the variable, method of measurement, and data source (either observations from game play in our system or the self-report survey administered to participants after RD1).

Variable	Method of Measurement	Data Source
2FA Decision	Enabled/Didn't Enable	System
Password Strength	Data-Driven Neural-Network Password Strength [Ur et al., 2017].	System
Signup & Login Times	Seconds each screen was in focus on participant's browser.	System
Internet Skill	Validated measure of web-use skills [Hargittai and Hsieh, 2012].	Survey
Security Behavior Intention	Validated measure of security behavior intention [Egelman and Peer, 2015].	Survey
Gender	Male/Female/Other	Survey
Education	HS or less/Some college/Bachelors or above	Survey
Age	Numeric	Survey

Table 5. Variables measured in our experiment.

12 COMPUTATION OF MTURK WAGE

Here we outline our computation of the average MTurk wage for the population. The figures in the computation were drawn from a Pew Research survey of Amazon Mechanical Turk workers, which provides wage estimates for different portions of the MTurk population [Hitlin, 2016].

$$\begin{aligned}wage_{mturk} &= 0.01(\$12) + 0.03(\$10) + 0.04(\$8) \\ &\quad + 0.07(\$7) + 0.11(\$6) + 0.2(\$5) + 0.52(\$4) \\ &= \$4.97/hr\end{aligned}$$

13 REGRESSION MODELS AND TABLES

13.1 Models Constructed

Table 6 summarizes the regression models that we constructed during our analysis and references the tables in which the results of these models are contained.

Outcome Var	Input Vars	Result Table
R1 2FA Decision	S, H, P, H:S, P:S, age, gender, education, IS, pwd. strength, SeBIS	Section 5 Table 1
R2 2FA Decision	R1 2FA Decision	Appendix Table 7
R2 2FA Decision	R1 2FA Decision, R1 Login Time, R1 Signup Time, interactions between times and 2FA decision	Appendix Table 8
R2 2FA Decision	R1 2FA Decision, R1 Login Time, R1 Signup Time, interactions between times and 2FA decision, S, H, P, P:S	Section 5 Table 2
R1 2FA Rationality	S, H, P, H:S, P:S, age, gender, education, IS, pwd. strength, SeBIS	Appendix Table 9
R2 2FA Rationality	S, H, P, H:S, P:S, age, gender, education, IS, pwd. strength, SeBIS	Appendix Table 10

Table 6. Regression models constructed. S stands for setting (Earn or Endowment, with Earn as baseline); H is a numeric variable for Hack percentage; P is a numeric variable for protect percentage; gender is a boolean variable with Male as the baseline; education is a three-level categorical variable with HS education or less as the baseline; decisions in prior rounds are boolean with False as the baseline; and IS (internet skill), SeBIS (security behavior intention), age, time measurements, and pwd. (password) strength are numerics. Input vars shown in table for R1 2FA Decision are those that were included prior to backward selection, only S, H, P, and P:S were retained after model selection.

13.2 Results Tables

The tables below present full regression model results for all findings presented in the text. We report the variable (one of those from Appendix 11), the log-adjusted odds ratio for the coefficient of that variable (O.R.), the 95% confidence interval (C.I.) for that O.R., and the p-value (where the significance threshold is 0.05).

Variable	O.R.	C.I.	p-value
R1 2FA Decision	83.11	[21.17, 326.24]	<0.001*

Table 7. Logistic regression results for 2FA behavior in RD2 modeled as a function of RD1 (R1) 2FA decision. McFadden Pseudo- $R^2=0.353$.

Variable	O.R.	C.I.	p-value
R1 2FA Decision	94.81	[22.04, 407.8]	<0.001*
R1 Login Times	0.94	[0.83, 0.97]	0.037*
R1 Signup Time	0.92	[0.73, 1.16]	0.484
R1 2FA:Login Times	1.10	[0.82, 1.47]	0.514
R1 2FA:Signup Time	2.20	[0.32, 15.26]	0.423

Table 8. Logistic regression results for 2FA behavior in RD2 modeled as a function of RD1 (R1) 2FA decision and RD1 costs (mean login time and signup time in RD1, both time measures do not include time spent on 2FA screens if 2FA was enabled). McFadden Pseudo- $R^2=0.522$.

Variable	O.R.	C.I.	p-value
Endowment	1.25	[1.05, 1.49]	0.015*
IS	1.15	[1.03, 1.29]	0.015*
<i>H</i>	15.38	[10.89, 21.73]	<0.001*

Table 9. Binomial logistic regression model with whether the participant made a utility-optimal decision in RD1 as the output factor and internet skill score, condition, and hack percentage as inputs, after backward selection. McFadden Pseudo- $R^2 = 0.141$.

Variable	O.R.	C.I.	p-value
Endowment	0.17	[0.01, 4.51]	0.293
<i>H</i>	1.21	[1.01, 1.87]	0.041*
<i>P</i>	1.03	[0.99, 1.90]	0.109
SeBIS	3.89	[1.12, 13.46]	0.039*
Endowment: <i>H</i>	1.02	[1.01,1.14]	0.09
Endowment: <i>P</i>	76.49	[0.41, 4162.32]	0.103

Table 10. Binomial logistic regression model with whether the participant made a utility-optimal decision in RD2 as the output factor and SeBIS score, condition, hack percentage, and protect percentage (as well as interactions between the latter three factors) as inputs, after backward selection. McFadden Pseudo- $R^2=0.078$.

Thought-Experiment Computations

Login Freq.	Income (wk)		All (\$741)		Most (\$554)		Half (\$522)		Little (\$261)		Very Little (\$52)	
	Risk	Utility	2FA	Risk	Utility	2FA	Risk	Utility	2FA	Risk	Utility	2FA
Daily	1%	\$0.07	✗	1%	\$0.05	✗	1%	\$0.025	✗	1%	\$0.025	✗
	1.5%	\$0.11	✗	2.0%	\$0.11	✗	2.2%	\$0.11	✗	4.2%	\$0.11	✗
	2.7%	\$0.19	✗	3.5%	\$0.19	✗	3.7%	\$0.19	✗	7.4%	\$0.19	✗
	4.9%	\$0.35	✓	6.5%	\$0.35	✓	6.9%	\$0.35	✓	14%	\$0.35	✓
	20%	\$1.42	✓	20%	\$1.07	✓	20%	\$1.00	✓	20%	\$0.50	✓
3× a week	Risk	Utility	2FA	Risk	Utility	2FA	Risk	Utility	2FA	Risk	Utility	2FA
	1%	\$0.07	✗	1%	\$0.05	✗	1%	\$0.025	✗	1%	\$0.025	✗
	1.5%	\$0.11	✗	2.0%	\$0.11	✗	2.2%	\$0.11	✗	4.2%	\$0.11	✗
	2.7%	\$0.19	✓	3.5%	\$0.19	✓	3.7%	\$0.19	✓	7.4%	\$0.19	✓
	4.9%	\$0.36	✓	6.5%	\$0.35	✓	6.9%	\$0.35	✓	14%	\$0.35	✓
20%	\$1.42	✓	20%	\$1.07	✓	20%	\$1.00	✓	20%	\$0.50	✓	
Weekly	Risk	Utility	2FA	Risk	Utility	2FA	Risk	Utility	2FA	Risk	Utility	2FA
	1%	\$0.07	✗	1%	\$0.05	✗	1%	\$0.025	✗	1%	\$0.025	✗
	1.5%	\$0.11	✓	2.0%	\$0.11	✓	2.2%	\$0.11	✓	4.2%	\$0.11	✓
	2.7%	\$0.19	✓	3.5%	\$0.19	✓	3.7%	\$0.19	✓	7.4%	\$0.19	✓
	4.9%	\$0.36	✓	6.5%	\$0.35	✓	6.9%	\$0.35	✓	14%	\$0.35	✓
20%	\$1.42	✓	20%	\$1.07	✓	20%	\$1.00	✓	20%	\$0.50	✓	

Table 11. Utility-optimal 2FA choices for different, hypothetical, categories of MTurk users. ✓ indicates that 2FA is utility-optimal given costs from our study, while ✗ indicates that 2FA is not utility optimal. We assume that 2FA offers protection from hacking 50% of the time.